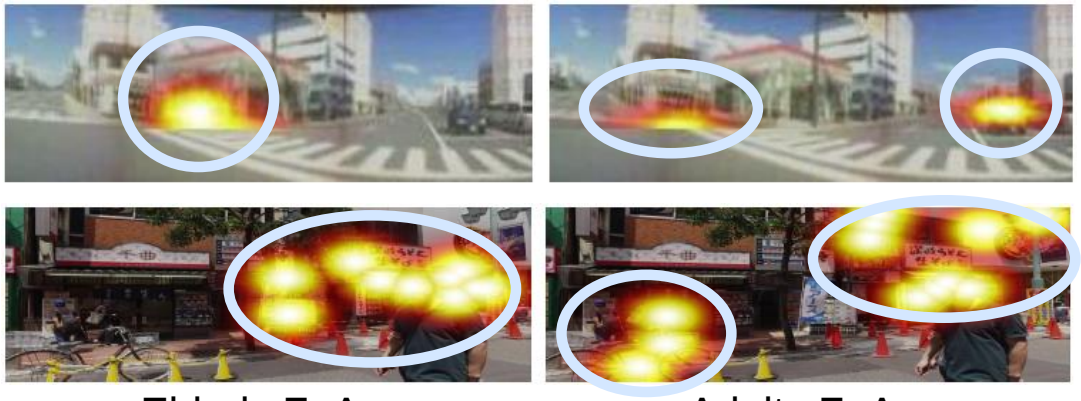
Translating Adult's Focus of Attention to Elderly's

Onkar Krishna Go Irie Takahito Kawanishi Kunio Kashino Kiyoharu Aizawa

NTT Corporation The University of Tokyo

Motivation

- FoA is a region of an image/video which attracts our attention
- FoA of elderly is significantly different from other age-groups **due to aging!**
- Existing FoA prediction models **fail to predict elderly FoA** as they are developed and evaluated on adults' eye-gaze data.



Elderly FoA Adults FoA

Our goal is to propose an approach to predicting the elderly FoA
for assisting their daily activities, such as driving, walking, and searching.

Challenges and Approach

- **Straightforward Approach** – training an FoA prediction model on elderly's data.

  ☹ Collecting a sufficient amount of data from elderly's is more challenging than adult, due to their physical or health conditions.

- **Assumption**: Correlation between adult's and elderly's FoAs can be characterized by the scene they are viewing.
- **Our Proposal**: **Image-to-image translation from adult's FoA to elderly's.**

  ☺ Leveraging well-trained models for adult's FoA for data efficient training.

Problem Setting

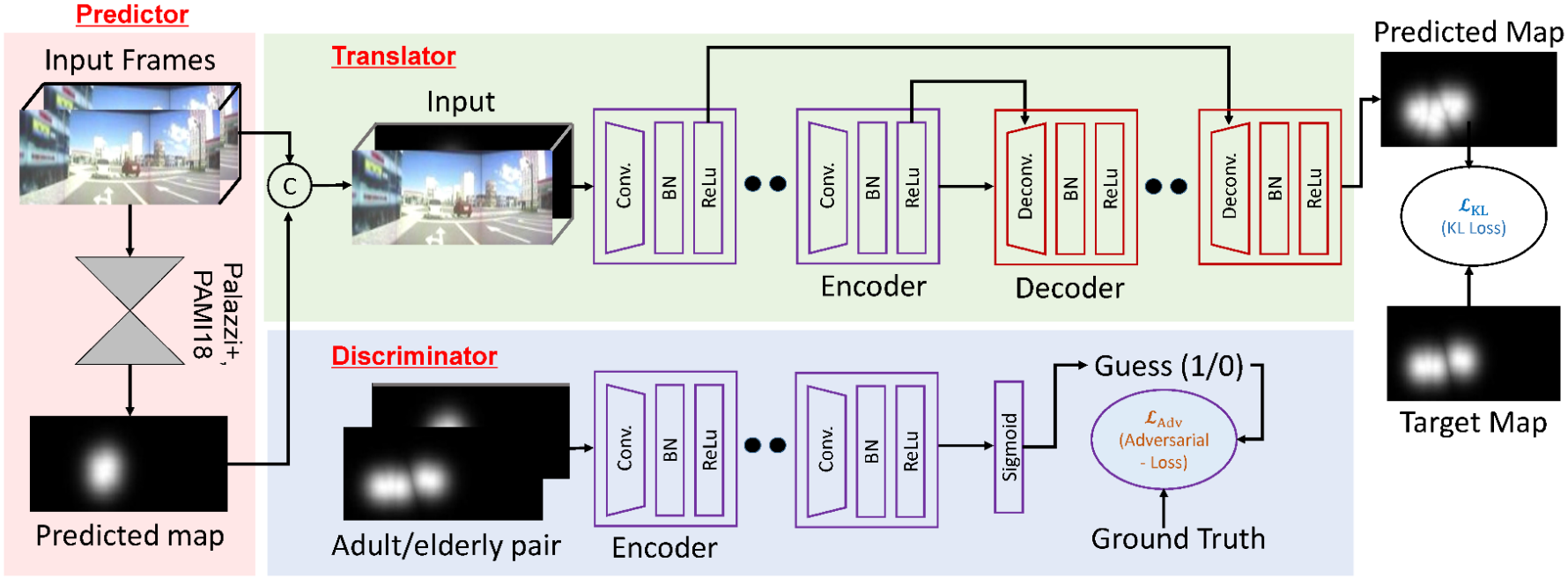
- **Input:** Sequence of video frames
- **Output:** Eldery's FoA maps for input videos viewed in two different scenarios

Driving
car driving scenario

Street Video
free viewing scenario during street walking

Network Architecture

Our model has **predictor network** and **translator network**
trained w/ support of auxiliary **discriminator network**



Loss Function Design

Our model is trained by minimizing both **Kullback-Leibler divergence** and **Adversarial loss**

$$\min_{D} \max_{T} \mathcal{L}_{\mathrm{Adv}}(T, D) - \gamma \mathcal{L}_{\mathrm{KL}}(T)$$

$\mathcal{L}_{\mathrm{KL}}$ loss requires the translator network *T* to output the ground truth FoA

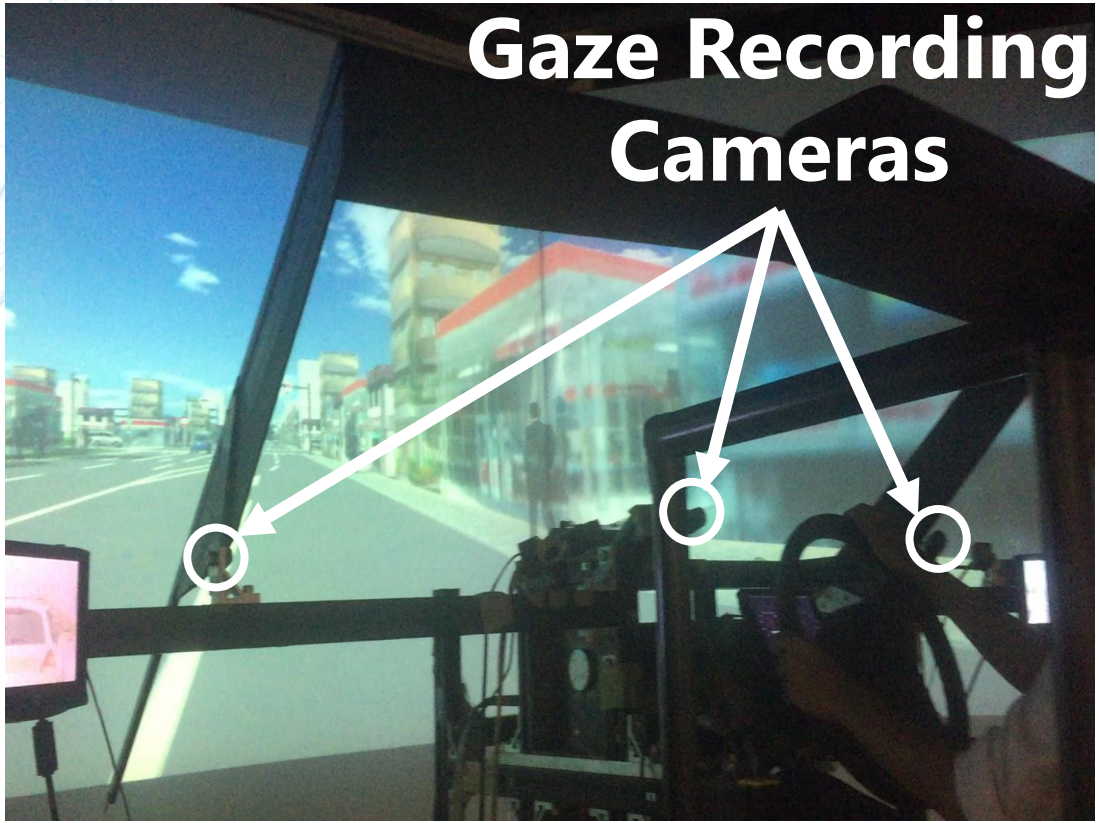
$$\mathcal{L}_{\mathrm{KL}}(T) = \sum_{n} \sum_{i} e^{*}_{n,i} \left(\log\left(e^{*}_{n,i}\right) - \log\left(e_{n,i}\right) \right)$$

$\mathcal{L}_{Adv}$ loss approximate the joint probability distribution of adult and elderly FoA

Dataset Construction

- We construct **Driving dataset and Street Video dataset** covering both adults and elderly.
- **Participants**
  - 18 participants (adults and elderly)
  - Adult's mean age **26 years**, elderly mean age **75 years**.
- **Eye-gaze for Driving**
  - Collected fixations of each participant while driving on a car simulator in real time
  - 9,713 FoA maps correspond to the 9,713 frames (train-test split 7,716/1,997)
- **Eye-gaze for Street Video**
  - Collected fixations of each participant while watching street-walking videos displayed on a monitor
  - 4,425 FoA maps correspond to the 4,425 frames (train-test split 3,532/893)

Driving Scenario



Result on Driving Dataset

Algorithm	CC ↑	SIM ↑	KL-div. ↓	Time (sec.) ↓
[Wang+, 19]	0.13	0.22	5.60	6.31
[Wang+, 15]	0.09	0.26	4.90	6.43
[Cornia+, 16]	0.26	0.42	9.97	2.71
[Palazzi+, 18]	0.64	0.53	4.06	7.48
[Palazzi+, 18] (fine tuned)	0.66	0.55	3.89	7.48
Ours	**0.91**	**0.79**	**0.80**	7.56

Ours achieves huge performance gain with this slight expense of run time compared to the base method

Result on Street Video Dataset

Algorithm	CC ↑	SIM ↑	KL-div. ↓	Time (sec.) ↓
[Harel+, 07]	0.24	0.49	0.82	4.10
[Itti+, 2000]	0.22	0.47	1.00	6.33
[Jiang+, 18]	0.27	0.46	2.21	9.23
[Cornia+, 18]	0.27	0.47	1.36	**2.74**
[Cornia+, 18] (fine tuned)	0.58	0.57	**0.58**	**2.74**
Ours	**0.72**	**0.71**	0.94	2.93

Ours outperform all the baselines for CC and SIM scores with slight expense of run time

Qualitative Results

NTT

Driving Dataset

Scene | Ground-truth | Ours | Palazzi+, 18

Street Video Dataset

Scene | Ground-truth | Ours | Cornia+, 16

Ours model accurately mimics the gaze of elderly people!

Summary

We introduced a deep image translation framework for predicting the elderly's FoA.

- *Accuracy:* our model can accurately mimic the elderly FoA while driving and street walking which can be useful in assisting elderly.
- *Novel Training:* adversarial training together with KL-divergence loss allows us to reach state-of-the art performance.